国外 Data Commons 平台的调查和分析

■ 吴雅威　张向先
吉林大学管理学院　长春 130022

摘要：[**目的/意义**]调研和分析国外 Data Commons（数据共享空间）的数据管理模式，为建设我国的数据共享空间提供借鉴。[**方法/过程**]通过梳理、归纳国内外数据共享空间的发展现状，对比和分析二者之间差距，并以美国 INRG 数据共享空间为例，从原则与协议、数据库与用户接口以及数据标识与关联等方面剖析其数据空间管理模式，为我国数据共享空间的建设及发展提出策略。[**结果/结论**]结合案例和我国数据共享平台现状，从总体规划、建设目标、要解决的问题、DC 总体架构和用户服务等方面提出具体建议。

关键词：数据共享空间　数据管理　数据服务
分类号：G250
DOI：10.13266/j.issn.0252－3116.2019.18.016

1 引言

数据时代的强大引擎持续推动着科技和社会向前发展，而"数据是新的燃料"也诠释了数据重要的资源价值[1]。不论对个人、组织还是国家，数据都是待挖掘的宝贵资产，未来最成功的组织可能就是那些能够利用数据资源实现有形或无形资产最大化的组织[2]。随着数据越来越有价值，科学研究和决策的数据获取、数据分析、数据共享和数据应用也变得极为重要，已得到越来越多国家和组织的重视。近年来，我国也已陆续开展数据管理工程建设，如 2002 年，科技部实施"国家科学数据共享工程"[3]，开展了不同领域的数据共享建设；2015 年，国务院印发《促进大数据发展行动纲要》[4]，提出我国的数据共享战略等。但发展至今，不同领域的团队或研究者仍面临着数据获取、数据分析和数据共享等问题，在许多研究项目中，因数据共享方面存在的隐私、产权和管理等问题，使数据获取和使用过程繁琐冗长，花费了研究者大量的时间和精力，阻碍了研究者将数据更全面地应用于实际问题的解决。因此，需要一个高效的、可持续使用的数据管理模式，提高研究人员获取、分析和共享数据的效率。作为一个庞大的、可互操作的数据共享平台，国外数据共享空间（Data Commons，DC）为科研人员和其他用户提供了新的研究模式，使异构数据源、分析方法和第三方应用得以融合，可显著提升和扩大研究人员或团体科学数据发现的速度和范围，进而提高个人和团队科研生产力。因此，我国有必要引入国外 DC 数据管理模式，并开展相应的 DC 数据服务。

2 Data Commons 的内涵、特点及理论研究进展

国外对于 DC 的研究和建设起源于 20 世纪 70 年代前后，最先应用于医学领域，而后延伸到科研、经济和社会政策等其他领域，主要用于解决各个领域突出的数据管理问题。多年来关于 DC 的理论研究不断充实与完善，包括 DC 规划、管理、运营、发展及相关法规等范畴，丰富了 DC 内涵的同时也保证了其特色性。

2.1 Data Commons 的内涵

在国外的数据科学领域中，DC 是一个对数据进行定位、储存和分析的网络基础设施，更是一个利于研究团体使用通用方法和工具分析与共享数据的数据共享空间[5]。以医疗领域中的基因组数据共享空间 GDC（Genomic Data Commons）[6]为例，GDC 是一个在肿瘤医学领域中促进精确分析的数据共享平台，它不但是一个数据库或工具，还是一个可扩展的知识网络，用以支持来自各类癌症研究计划的基因组和临床数据的导入、标准化与最优化利用。因此，本文认为 DC 的内涵是：以科学及用户服务为目标，严格遵循数据法规，为各类数据用户解决数据管理问题，将数据获取、数据聚

作者简介：吴雅威（ORCID：0000－0001－9703－8731），博士研究生，E-mail：727888263@qq.com；张向先（ORCID：0000－0003－3186－2677），教授，博士生导师。
收稿日期：2019－01－02　修回日期：2019－03－31　本文起止页码：137－146　本文责任编辑：易飞

合、数据标识、数据分析、数据应用和数据共享功能集于一身的高效数据生态系统平台。

2.2 Data Commons 的特点

DC 在国外数据管理领域中应用广泛，在不断的发展和实践中也呈现出以下不同于其他数据管理模式的特点：

2.2.1 功能和作用 DC 注重数据集成、数据关联、数据发现、数据审查、数据分析、数据应用、数据共享等功能的一体化建设，进而打造出专业化的数据生态系统，致力于打破数据垄断，增强科研人员对数据的复用和创新，从而挖掘数据的深层价值。

2.2.2 数据处理方式 DC 在处理不同数据时，遵循一整套统一的标准化流程，如：①前期处理，包括数据的录入、审查和筛选等，并通过元数据或数据字典将数据进行标准化、一致化处理；②中期处理，包括根据用户需要进行数据分析、可视化、透明化和匿名化等；③后期处理，包括数据共享、数据出版、数据关联和数据推送等。

2.2.3 用户服务 DC 主要以实体空间和虚拟平台两种形式为用户提供服务，实体空间以深度接触、交流和挖掘用户需求为主；而在虚拟平台中，DC 除提供必需的数据管理和分析方法外，还可为用户开发和研究自己的数据分析方法、工具提供培训和指导，在提高数据可访问性和可互操作性的同时，也可提升用户的数据素养和数据意识。

2.3 Data Commons 理论研究进展

2.3.1 国外数据共享空间理论研究 国外对 DC 展开的理论研究较早，也较为完善，总体可归结为 3 个方面：

（1）DC 的设计和实施。近年来，国外对 DC 设计和实施的研究取得了新成果，包括 DC 框架设计、解决资金问题和在各领域的应用等。DC 构建方面，F. Molinari 等[7]、J. Mansell 等[8]皆设计了 DC 的一整套蓝图并提出 DC 是将来可替代其他数据管理模式的高信任度、低成本的数据共享平台；关于资金问题，R. L. Grossman[9]就科学研究领域的 DC 资金问题提出了指导建议；针对 DC 的应用，S. P. French 等[10]提出应针对社会化大数据创建 DC；美国癌症研究机构[11]和 S. L. Volchenboum 等[12]都为能深入癌症研究而支持创建儿童疾病数据的 DC。

（2）DC 的管理和运营。在 DC 的管理、运营方面，国外学者分析了 DC 当下要解决的问题，如 S. A. Sansone 等[13]提出 DC 数据的标准化及共享原则；C. Bizer 等[14]提出 DC 数据的元数据标准化、RDFa 和微格式的解决方案；N. Purtova[15]明确了 DC 的共享边界和社会困境；此外学者还研究了如何改善用户界面及管理数据，如 M. Morgan 等[16]提出改善 DC 与用户沟通的接口，以更好地与用户交流；Z. Su 等[17]就 DC 如何改善疾病数据管理的途径提出了建议；C Scott 等[18]提出了如何有效分析 DC 中的数据集问题等。

（3）DC 的相关法规。P. N. Halphin 等[19]研究了 DC 在运营和管理过程中要遵循的法规，指出 DC 要持续发展必须依靠相关法律；OCLC[20]提出政府和相关机构应制定相关政策，以此来促进用户获取、分析和共享 DC 数据，鼓励用户在 DC 上管理个人数据库，并基于法规提供具备用户属性的数据；J. Yakowitz[21]讨论了数据共享受到阻碍而导致的悲剧，提出数据共享要注意隐私安全法，最大程度规避因隐私和产权产生的不良后果。

2.3.2 国内数据共享空间理论研究 国内对于数据共享平台的理论探讨可归纳为两方面：

（1）借鉴并引入国外理论成果经验。对国外理论成果的引入和借鉴，可归结为：①内部建设策略，包括平台建设和服务模式、功能与特点、目标与内容、人员数据素养 4 类；②外部条件支持，包括政策和法规、经费来源、机构合作、发展与局限性 4 类。如宋秀芬等[22]剖析了国外 3 所著名大学的数据平台特征、功能及局限性，提出我国应从平台功能、政策支持、数据标准、教育培训与合作交流等方面进行建设；覃丹[23]、完颜邓邓[24]都对英美两国高校的数据共享平台进行了分析，从平台引进、政策制定、数据服务细分、资金来源、建设模式等方面给出发展建议；杨鹤林[25]、殷沈琴[26]分别介绍和评估了国外数据平台的模型、思路、特色、进阶功能、元数据标准、在线分析功能等。

（2）我国数据共享空间的建设和发展研究。其内容主要为两方面：一是自上而下式分析，即对已建成的典型数据平台的功能、特点和服务等进行评析，如朱玲等[27]论述了北京大学开放数据平台的构建过程；殷沈琴等[28]评析了复旦大学数据平台的系统选型与功能；二是自下而上式设计，即从平台建设体系、服务体系、评价体系、数据管理和政策制定等基础条件提出建议。如邓仲华等[29]从保障信息安全、拓展服务内容、营造共享氛围入手设计了“互联网 + ”环境下数据共享平台的建设模型；刘兹恒等[30]提出基于学科服务平台或机构知识库来建设科研数据共享平台、制定共享政策、提升数据素养；刘桂锋等[31]从平台建设基础、数据、管

理功能及效果与影响 4 个方面构建了数据平台评价指标体系。

可见,国内外在对数据共享空间的理论研究中各有侧重,也由此体现出二者数据共享平台设计和实施、管理及运营等理论研究方面的差距:

(1)在设计和实施理论中,国外研究最注重数据共享空间一整套蓝图的前期设计和准备,尤其关注资金问题、框架设计以及明确数据共享空间后期可以用来解决什么问题、如何解决等;在我国,由于数据共享空间的建设和发展较晚,其理论研究主要是在借鉴英美等国家共享空间建设经验的同时探索适应本身状况的共享空间的实施和运营框架,从而采取边探索、边设计、边实践的策略。

(2)在管理和运营理论中,国外侧重于先对数据共享空间的困境和问题进行分析并解决,其次是对共享空间中数据和用户服务的管理与完善,对于数据集,要求遵循标准化及共享原则,使用元数据、RDFa、微格式等进行数据标准化,从而满足用户各种数据应用的需求,属优化阶段;我国现阶段研究侧重于数据共享空间的系统选型、基础建设、改善服务、评价体系建设、政策制定以及探索共享空间管理和运营的具体路径,虽已有典型数据空间如复旦大学、中国科学院等少数机构的数据共享空间投入运营,但仍处于建设和发展阶段。

3 国内外 Data Commons 的实践发展

3.1 国外数据共享空间的实践发展

国外 DC 经过几十年来的建设、发展和完善,已在各个领域和国家得到应用,其中美国建设的时间较早,现已逐渐传播至英国、澳大利亚等国家,并建成有各国特色的 DC 管理模式。通过调研,本文从研究领域、功能特色和运营模式等方面总结了国外 8 家典型 DC 的数据管理模式,如表 1 所示:

表 1 国外数据共享空间的实践和发展

国外 DC 名录	所属国家/机构	研究领域	功能/特色	管理/运营模式	网站/资源链接
Social Change Data Commons(SCDC)	美国公共图书馆	社会问题公共政策	激励公民参与,吸收新观点,连接社会与数据,以解决问题和影响政治决策为核心理念	网站平台+博客+物理空间	https://www.calfund.org/social-change-data-commons/
Data Commons(DC)	新加坡	科学领域	支持科学研究,促进数据传播和发现,开发分析工具,匿名化数据,创建用户数据生态系统等	物理空间+网站平台	https://datacommons.nus.edu.sg/
Australian Research Data Commons (ARDC)	澳大利亚国立大学	科学领域	创建 DC,供研究界使用,以共享一系列学科 FAIR 型数据,满足数据密集型、跨学科和全球协作研究需求	物理空间+网站平台	https://ardc.edu.au/planning/events/top-10-fair-data-things-global-sprint
NIH Data Commons (NIH-DC)	美国国立卫生研究院	医学领域	使研究人员获取、互操作和可重用数据来加速数据发现。用创造性新方法组合、分析和提出新问题,以产生新的知识等	物理空间+网站平台+实体机构	http://www.bio-itworld.com/2017/11/07/nih-launches-data-commons-pilot-with-9-projects
P2P Data Commons (P2PDC)	瑞士	商业领域	实施数据对等协议,数据管理个性化,以安全认证口令牌、API 和算法控制数据用户且具有数据可移植性等	物理空间+网站平台	https://www.tokencommons.org/
Data Commons(DC)	新西兰	科学领域	一种基于信任和协议的奖励和鼓励数据集成、重用、共享的数据生态系统等	物理空间+网站平台	http://datacommons.org.nz,2017.
Data Commons(DC)	美国斯坦福大学	教育领域	研究和开发分析方法、算法和软件并将其应用于数据分析,提高数据的可访问性等	物理空间+网站平台	https://sdsi.stanford.edu/data-commons
INRG Data Commons (INRG-DC)	美国	医学领域	创建较完整的管理和运营架构,包括:共享系统、数据生命周期流程、运营环境、数据审查分析等	物理空间+虚拟空间	http://europepmc.org/abstract/MED/28561664

由表 1 可见,DC 的实践领域涵盖基础科学、医疗卫生、公共服务和教育等领域,在发现社会潜在问题、挖掘公民需求、影响公共政策和提供决策等方面做出了重要的贡献。

(1)功能与特色方面。DC 具备为研究人员获取、分析、监护、应用和共享数据的能力,在基于相关数据政策的前提下,发挥着服务社会与公众的作用。经过多年发展和完善,DC 的数据管理模式也正发生转变,不仅包括物理空间、实体、微博和虚拟平台的项目,还不同程度上担当着智库的角色,有逐渐发展成为特色智库的趋势,通过对数据的管理分析为研究者和政府提供决策依据,使得用户做出更明智的决策。

(2)数据管理与分析方面。DC 作为一种新的工具、数据库的扩展或一类网络基础设施,降低了用户数据获取的复杂性和成本、提高了数据分析质量、简化了数据应用过程等,如利用图形化展现技术实现数据的

可视化分析、链路分析、血缘分析和影响分析,实现系统间应用集成关系的可视化展现,为用户提供多层级、细粒度的分析结果展现,等等。此外,DC打破了数据获取和分析的局限,利用常见基础设施分析、共享数据,为科研团体提供了可互操作的平台,如美国某非盈利公司[32]开发了DC云计算基础设施来支持科学研究,如开放科学数据共享云等,用户包括大学、非营利组织、公司和政府机构等。

(3)管理和运营模式方面。大多数DC在建设初期会对其设计、系统选型、功能和服务规划以及相关规则进行严格控制,如设计一整套执行计划、制定数据管理生命周期流程和相关技术规则与协议等,尤其通过用户激励政策,鼓励各类用户共同参与DC的管理和运营。由于前期准备工作充分,后期建成的网站平台、虚拟空间甚至实体公司功能往往能更加完备,服务更多样化,此外,DC这些功能都能在用户界面中体现,研究人员能够与这些平台进行互操作,以REST-API编程方式与DC建立接口,实现查询和下载数据,驱动当前数据门户,如为项目、文件和案例创建DC数据模型的索引视图、结果分析图、共享路径图及收集相关信息等,因此,DC尤为重视用户界面的建设和完善。

3.2 我国数据共享空间的实践发展

我国近十几年来正在建设和发展数据共享空间,较早的如清华大学中国经济社会数据中心于2009年开始投入运营,近年来如复旦大学和北京航空航天大学的数据空间都已初具规模。如表2所示,本文选取我国8家具有代表性的数据共享平台,通过对数据源涵盖领域、平台功能与服务、合作机构等方面的调研和分析,总结和归纳数据共享平台的理念和目标、功能和服务及建设和管理机制等。

表2 国内典型数据共享平台的发展现状

数据平台名录	隶属机构/运营时间	数据源领域	平台功能与服务	合作机构
复旦大学社会科学数据研究中心[33]	复旦大学/2013年	国内外科学研究数据等	数据监护、数据共享、数据引证、数据分析等	哈佛大学 data verse
北京航空航天大学数据共享平台[34]	北京航空航天大学/2014年	主要为内部数据库等	数据审计、交换与数据共享、数据应用等	校内各部门
国家人口与健康数据共享平台[35]	基础科学数据中心/2016年	人口与健康数据等	数据集成、共享等	清华大学、北京大学等
国家基础科学数据共享服务平台[36]	中国科学院计算机网络信息中心等/2013年	基础科学领域数据资源等	数据标准化、数据发现、检索、下载等	中国科学院、国内高校和其他科研院所
国家地球系统科学数据共享平台[37]	中国科学院地理科学与资源研究所/2011年	环境、区域等地球科学数据等	数据查找、数据下载、数据整合与共享等	中国科学院地理科学与资源研究所等国内外40家机构
武汉大学高校科研数据共享平台[38]	武汉大学/2012年	物种资源数据库、读者调查数据等	数据收集、存储、快速检索、共享、再利用等	校图书馆、开源软件 Dspace 等
华中科技大学社会科学数据中心[39]	华中科技大学/2012年	电子科技、系统工程数据等	数据管理、数据挖掘、决策支持、系统设计开发服务等	华中科技大学、中国高校社会科学数据中心、电子信息与通信学院等
清华大学中国经济社会数据中心[40]	清华大学/2009年	以经济调查数据和宏观截面数据为主体等	数据收集、数据处理等功能、经济调查和政策发展研究等	清华大学经济管理学院、人文社会科学学院等

由表2可见:①理念和目标。我国数据共享空间的建设理念围绕着用海量数据为其附属机构或社会公众提供数据资源、数据服务、决策支持和软件开发等,其主要目标是实现对数据的存储和管理,如华中科技大学科学数据中心所提供的软件设计和开放服务。②建设机制。大致分为两个方面:一是合作共建,这是大多数机构都采用的方式,如复旦大学社会科学数据研究中心与哈佛大学 dataverse 合作建设数据中心、国家地球系统数据平台等;二是根据本身机构特色自主建设,如北京航空航天大学数据共享平台采取校内机构合作方式建设。③功能与服务。对于功能方面,我国数据共享空间的功能多集中于数据集成、存储、标准化、检索、分析、共享等方面,基本涵盖了数据生命周期的主要环节;用户服务方面,我国数据共享空间主要面向高校、研究机构和政府等用户,已建成的部分数据共享平台主要为用户设置了如用户注册、登录、访问、下载和数据分析等基本服务,少数共享平台为用户提供应用、协作研究和决策支持等增值服务。

3.3 国内外数据共享平台的对比分析

本文选取国内外各5家典型数据共享空间,分别从平台管理和运营、平台功能和用户服务3个方面进行对比,如表3所示,发现与国外相比,我国数据共享空间建设仍存在差距和不足。

(1)平台管理和运营。与国外相比,我国数据平台的管理和运营较为薄弱,缺乏一整套的数据治理规划、数据生命周期流程、合理的管理结构和各项协议设

吴雅威,张向先. 国外 Data Commons 平台的调查和分析[J]. 图书情报工作,2019,63(18):137 – 146.

表 3　国内外典型数据共享平台的功能与服务对比

国内外平台		平台管理和运营	平台功能		用户服务	
			数据处理前期	数据处理后期	基础服务	增值服务
		数据治理规划、完善数据生命周期流程、合理管理结构、各项协议设定等	数据安全认证、提交、审查、存储、标准化、检索、分析、可视化等	下载、重复分析、再次利用、共享、发布、出版等	管理数据、个性设置、连接用户和数据、用户培训等	协作研究、鼓励用户参与、决策咨询等
国外	INRG Data Commons	√	√	√	√	√
	Australian Research Data Commons		√	√	√	√
	P2P Data Commons	√	√		√	
	Data Commons (新加坡)	√	√	√		√
	Social Change Data Commons	√	√	√		√
国内	复旦大学社会科学数据研究中心		√	√	√	
	北京航空航天大学数据共享平台		√	√		
	国家人口与健康数据共享平台	√		√	√	
	国家地球系统科学数据共享平台		√			
	华中科技大学社会科学数据中心		√		√	

定等,如除复旦大学社会科学数据平台的元数据 DDI 标准外,其他平台均无明确说明,直接影响了数据空间的实际应用,导致仅少部分功能可提供用户使用,实用性不强。在总体规划和政策协议方面,国内目前虽有对数据管理的宏观政策制定和布局,但缺乏中观和微观层面的管理规范来引导和激励数据管理的发展,如平台构建要素、使用和评估规范、人员激励政策、用户规范和必要强制性措施等,从而使数据管理进程缓慢,影响了实施效果。

(2)平台功能。国内多数数据共享平台在数据处理的前期和后期功能设置较为完善,主要为数据集成、数据存储、数据分析和数据共享等功能,但在实际应用和操作中未对用户开放,有些功能还处于封闭中或有待改善,如多数平台功能只包括:数据检索(有些检索方式单一、高级检索项缺失等)、导航(存在部分空链接、无链接等)、下载(只少量数据允许下载,或需提交申请,步骤繁琐)、分析及可视化(在线数据分析和可视化开放程度低,导致用户利用率降低)和共享等,影响了用户的体验感及对数据平台的评价。

(3)用户服务。国内多数数据共享平台在服务方面的规划和建设呈现出重功能而轻服务的现状,对于用户基础服务仍未完全实现,如访问数据、鼓励用户参与、个性化设置、用户培训等仍有待完善;增值服务方面的不足较为突出,如协作研究、连接用户和社会、决策咨询等比国外少,服务方式单一、低效,仅有少数提供决策支持和用户培训等服务,如复旦大学社会科学数据平台、华中科技大学社会科学数据中心等。导致上述问题的原因可归结为:总体管理和运营规划缺失、用户需求开发不足,缺乏概念推广、需求调研、服务多元化建设等系统化的保障,导致用户关注度不高,缺乏信任感和用户互动性,从而难以展开全面、彻底、有效的以用户为中心的服务。

4　国外 Data Commons 案例分析——以美国 INRG Data Commons 为例

4.1　INRG-DC 启动背景

在世界范围内,儿科癌症虽不多见,却仍是医疗领域的难题,因儿童癌症病例共享数据的缺乏而使得深入研究陷入瓶颈,而 DC 的出现为支持该科学研究找到了一种变革性的方法。2004 年,由北美、欧洲、澳大利亚和日本儿童癌症组织代表合作组建了国际神经母细胞瘤风险组(INRG),准备开启儿童癌症数据 DC,通过分析和共享数据找到最佳治疗办法[41]。

4.2　INRG-DC 管理模式

INRG 组建了专家团队,参照科学数据管理的

FAIR 原则,即可查找性(Findable)、可访问性(Accessible)、互操作性(Interoperable)和可重用性(Reusable)4个数据管理原则制定了 INRG-DC 的一套数据管理模式,其主要内容如下:

4.2.1 总体规划 INRG 制定了一整套设计、建设和管理 DC 的规划,如:①前期计划,包括资金、建设框架、管理和运营等;②中期计划,包括规范数据生命周期管理流程与完善数据管理体系等保证数据共享空间建设和发展的科学性、系统性、层次性和可持续性;③后期计划,主要是对平台功能和用户服务的指标性评价等,并作为改进的依据。

4.2.2 数据库与用户接口 芝加哥大学研究信息中心(CRI)研究人员设计并构建了 INRG 表型数据的数据库,并开启用户前端接口,基于协议,任何人都可以查询和使用相关数据。从 1980 年至今,数据库中已积累了 18 000 多位患者的数据并定期更新。除了基本数据,数据库还可以通过 API 过滤生物标本数据以预判可用性,大大提升了数据获取速度和准确度,简化了数据库与用户之间的交互过程,可直接连接到目标数据。

4.2.3 数据字典 基于相关标准和规则,INRG 建立了儿科癌症患者 DC 数据分类与分析系统和数据标准化系统,与每个地区的统计人员共同创建标准数据字典,将所有数据元素映射到该框架中,主要是将包含 1990–2002 年间全世界诊断出的 8 800 位患者临床数据进行标准化和同构化,并使数据得到充分利用,为数据分析、关联打下基础。

4.2.4 数据标识与关联 对于数据标识符问题,儿童肿瘤学组(COG)为每个数据分配通用样本标识符(USI)。USI 与随后生成的任何示例、数据和其他信息相关联,使得目标数据集中的所有样本都有 USI 链接回 INRG 数据库中的数据,直接关联数据集,保障了各类数据可被直接调用,并能够做到数据之间的关联,使得用户完成对数据较为全面的对比和分析,拓宽了使用数据的范畴。

4.2.5 数据审查和监护 DC 的操作中心是一个监护和审批程序,要求用户通过门户正式访问数据,数据访问由 DC 审查机制单独管理,可通过权限获取 INRG 临床数据和美国国立生物技术信息中心(NCBI)基因组数据,而后 DC 将临床数据存储为一个对象,先通过审查模式对数据质量和数量进行审查,而后启动虚拟机使用命令行工具进行数据分析,避免用户使用错误数据导致错误结果[42]。

4.2.6 原则及协议设定 INRG 调集相关专家制定出了 INRG-DC 的一套操作原则和协议。相关协议包括:①协同设计规则,要求 DC 与技术专家、科学家、用户和相关政府组织合作,交换 DC 实施的建议;②元数据规则,DC 需要元数据、词汇表控制及标准化数据库的各种数据元素,使数据容易被搜索、发现和关联;③匿名化与安全协议,DC 必须了解用户目的和需要,保障为用户分配共享资源的适配性和安全性,匿名化也可防止数据滥用等。

4.2.7 用户评价和需求分析 DC 通过内、外部数据的可视化分析工具和算法来完成对数据不同层级的分析,如数据应用频率可视化、相关图形可视化和诊断检测等,进而从不同层次分析数据的内、外部特征,用户还可通过思维导图认知数据处理过程,甚至可亲身操作基因组数据,如通过基因组 DC 应用程序接口从 DC 中提取数据,并与临床数据进行比对,完成后续分析和处理产生结果等过程,最后根据结果为用户提供最合理的决策和建议。上述过程既可探究和满足不同用户的需要,也可作为评价指标,在每个操作环节中得到用户反馈。

4.3 INRG-DC 建设成果

4.3.1 用户数据生态系统 INRG-DC 创建了一个本体化的,鼓励数据集成、数据分析、数据共享和数据应用的数据生态系统。该数据生态系统试图在最大化相关者利益的基础上,实现数据集成、聚合、分析和共享的完整数据生态周期模式,逐渐消除用户对于数据产权和数据隐私方面的忧虑,建立起用户对 DC 的信任感,进而促进从改变用户意识到促进用户行为的转变,以此为核心推动数据科学的进步和发展。

4.3.2 改善数据生命周期流程 通过 DC 加速了以联盟为主的协作发现、数据开发、数据归属、数据分析和数据共享的强大流程,增强了数据的互操作性和访问性。DC 基础架构、策略和流程的全面协作开发,最大好处之一是能够找到治疗儿科癌症的新型个性化医疗方法,进而识别最需要积极治疗的儿童群体,同时减低无效治疗的风险。

4.3.3 创设 DC 管理运营环境 基因组分析的发展以及数据存储和计算资源的民主化,为 DC 管理儿科癌症数据提供了理想的计算环境,实现了收集、标准化和聚合患儿不同表型、基因组数据和其他数据在 DC 内的互相连接。如今 DC 在可持续性运营环境下的运营不但促进了数据的应用,还对儿科癌症研究产生了积极影响,更为诊断和治疗患有肿瘤疾病的儿童提供

了新颖的方案。

可见,INRG-DC 在帮助解决医疗领域中的数据管理问题时担当了重要角色,这归功于 INRG-DC 良好的建设、管理和运营。从 DC 的总体规划到具体实施,不但形成了涵盖数据库与用户接口、数据标识与关联、数据审查监护和原则及协议设定等一系列数据功能与服务模式,还取得了包括建设数据生态系统和改善数据生命周期流程等成果,更具体、更微观地实现了集平台建设和用户服务的一体化管理。因此,我国在解决数据管理问题时应适当借鉴和引入 INRG-DC 的建设与管理模式,以设定总体规划、具体目标并针对主要问题为中心搭建和制定适应我国数据环境的 DC 架构和建设策略,为用户提供更高效、完善的服务。

5 我国 Data Commons 建设的总体框架及策略

5.1 DC 的总体规划、目标和要解决的问题

5.1.1 DC 总体规划　作为设计和建设数据共享空间的前期工作,总体规划及布局尤为重要,如制定一整套完备的前期设计蓝图(包括资金、建设框架、管理和运营等)和完整的数据生命周期流程等,通过完善数据管理体系来保障数据共享空间建设和发展的科学性、系统性、层次性和可持续性。

5.1.2 DC 的目标　我国 DC 建设的目标主要包括:①推进数据共享进程。作为数据共享空间,目的就是让数据实现最大限度地被管理并用于解决实际问题。②优化用户服务。通过提供以用户为中心的数据服务,克服用户以往遇到的数据管理障碍。③促进交流与合作。通过鼓励和促进数据间、用户间的交流和合作,充分发现并挖掘数据价值。④拓宽数据服务领域。DC 充当连接数据和用户的第三方角色,应基于数据法规,扩大服务范畴。

5.1.3 DC 要解决的问题　数据共享空间的建设与实施旨在帮助科研人员和其他用户解决将大量分散的且具潜在价值的多源异构数据集成化、标准化、价值最大化和分配最优化等问题。通过 DC 的有效管理,将不同领域的目标数据与数据用户快速连接,利用先进数据分析和管理技术实现数据、技术、人三者之间的高度结合,增强 DC 用户处理敏感、大规模和非结构化数据的能力,建设成从数据源确定、用户与数据交互、数据共享与应用到相关者利益合理分配等过程的完整、高效的数据生命周期模式,使得最具价值的数据能够通过 DC 这种有效途径被最合适的用户所使用,达到数据共享和高效应用的目的。

5.2 DC 的总体框架及建设策略

5.2.1 DC 总体架构　我国数据共享空间的总体架构(见图 1)应由用户层、用户接口层、应用与服务层和数据资源层 4 个部分构成。首先,各类数据用户通过数据共享空间的用户接口层将各领域数据资源提交、共享到 DC 内部;然后在应用与服务层,用户对目标数据进行管理,整个过程在 DC 管理和运营人员的监护下完成,保障了相关利益者层的利益均衡;最终,各层之间相互关联、相互协同及相互作用共同构成了 DC 完整管理机制的数据共享空间平台。

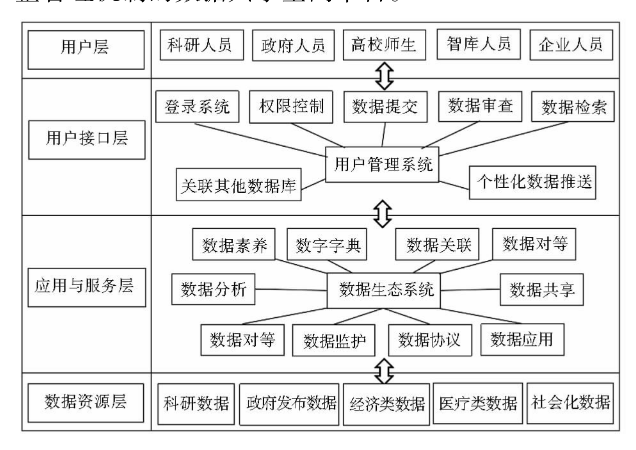


图 1　我国 DC 数据共享空间的总体架构图

(1)用户层。该层为 DC 运营中服务的主要对象和人群,根据 DC 在不同领域不同机构(包括高校、政府、企业、研究机构等)的建设和应用,其用户主要可分为科研人员、政府人员、高校师生、智库人员、企业人员和普通用户等,此外,DC 还应明确和制定各类利益群体的利益分配机制和原则。

(2)用户接口层。DC 接口层属于用户的基本操作层,主要通过网站、移动终端等接口负责直接与用户交互,处理用户请求和向用户提供各项服务,并协同用户完成数据管理和共享等一系列活动。用户接口层布局围绕着用户管理系统展开,主要包括登录系统、权限控制、个性化数据推送及与其他数据库关联,而用户管理系统又可与应用与服务层、数据资源层呈现映射关系且相互协调、反馈,即通过用户对数据分析、数据对等和数据共享等服务的利用来反馈用户接口的效果、性能以及数据资源的质量,反之亦然。

(3)应用与服务层。该层是 DC 的核心层,负责向用户提供各类数据应用与服务,围绕着数据生态系统功能展开,主要包括数据获取、数据字典、数据关联、数据分析、数据应用和数据共享等应用与服务。针对国

内数据管理的短板，该层以中观和微观层面的管理为重点，以多种应用与服务相融合的服务为核心的策略去引导和激励数据管理的发展，面向更广泛领域的用户，展开全面、彻底、有效的数据服务，实现精准识别和挖掘用户的现实和潜在需求，培养用户获取数据、分析数据和应用数据的数据素养和意识，消除以往用户接受数据服务的障碍。

（4）数据资源层。指 DC 管理的数据类型及数据来源，是数据管理服务的基础。数据种类包括自然与社会科学数据，如经济、医疗和社会化等领域数据；数据来源包括 3 类：①不同类型用户的提交与共享；②关联外部数据源（如企业、政府等发布的数据）；③DC 本身拥有的数据。确定数据来源后，可通过 DC 的数据字典、数字对象标识（DOI）或元数据对数据进行标准化、结构化，进而建立可持续、可分析的数据模型，支持与其他 DC 或云对等、连接，使数据更易利用、可互操作、网络容量级分析和共享等服务。

5.2.2 DC 的建设策略

（1）DC 的建设和运营。主要分为建设主体和运营责任主体，前者是指可以承担建设我国 DC 任务的机构或组织，如图书馆（以研究型图书馆和高校图书馆为主将传统图书馆的学术交流、咨询等服务拓展到数据阶段，使图书馆用户服务更具针对性、实用性[43]）、各研究机构的信息技术中心和数据管理机构等；后者包括软件提供商、政府人员、管理团队、技术团队、各类用户和其他利益相关者。此外应注意平衡用户和相关利益团体的利益分布。

（2）DC 的用户服务。分为完善基础服务、发展增值服务和用户激励措施 3 个方面，基础服务如访问权限、鼓励用户参与、个性化设置和用户培训等；增值服务如协作研究、连接用户和社会、决策咨询等；用户激励措施如降低数据获取成本、推送服务（特色工具和服务等）、宣传推广、奖励机制等，此外还应考虑解决安全限制，及时补充 DC 空间数据、工具和方法空白，促进业务的有效外展和公众参与等问题。

（3）DC 的相关协议。DC 责任主体应与相关机构制定数据空间和用户管理相关规则，如：①协同设计规则：DC 管理需依靠技术专家、科学家、用户和相关政府等协同管理；②元数据规则：DC 需完整有效的元数据、词汇表和数据命名规则，使数据被搜索、发现和关联；③控制协议：管理用户使用数据，可构建个人信息管理系统（PIMS）增强用户管理机制[44]；④透明度和匿名化协议：透明化管理用户需求，为用户安全、合理地分配共享资源，匿名化可防止数据滥用等。

6 结语

国外 DC 的理论和实践在数据管理领域中的较早应用体现出了 DC 的独特管理优势，其建设管理和用户个性化、自主化交互服务模式的特点可为我国建设和发展数据共享空间提供适当借鉴，进而用以解决我国数据管理方面的难题。鉴于我国数据共享平台功能和服务建设的不足及国内外数据环境的差异，本研究提出我国应从 DC 的总体规划、目标和要解决的问题中入手，创建适合我国数据环境的 DC 架构，同时注意区分 DC 建设和运营的相关责任主体，从平台建设、充分发挥功能、完善用户基础与增值服务和设定 DC 相关协议规则等方面建设和管理 DC，最终促使数据共享平台成为科研人员及其他用户管理和共享数据的重要渠道，从而使之更好地发挥科学数据管理和共享的职能和作用。

参考文献：

[1] GROSSMAN R L, HEATH A, MURPHY M, et al. A case for data commons: towards data science as a service[J]. Computing in science & engineering, 2016, 18(5): 10 – 20.
[2] 数据共享空间启动项目[EB/OL]. [2019 – 01 – 09]. http://www.bio-itworld.com/2017/11/07/nih-launches-data-commons-pilot-with-9-projects.
[3] 张先恩. 国家科学数据共享工程[J]. 科学中国人, 2004(9): 11 – 13.
[4] 国务院. 国务院关于印发促进大数据发展行动纲要的通知[EB/OL]. [2019 – 01 – 09]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
[5] National Institutes of Health. Newly launched genomic datacommons to facilitate datavand clinical information sharing[EB/OL]. [2019 – 01 – 09]. http://www.nih.gov/news-events/news-releases/newly-launched-genomic-data-commons-facilitate-data-clinical-information-sharing.
[6] 吴雅威, 魏来. 国外 Data Commons 的发展及其构建初探[J]. 情报资料工作, 2017, 38(6): 41 – 48.
[7] MOLINARI F, MORELLI N, TORNTOFT L K, et al. OpenDataLabs: new infrastructures for open datacommons[EB/OL]. [2019 – 01 – 09]. https://www.forskningsdatabasen.dk/en/catalog/2372370890.
[8] New Zealand Project. Data commons blueprint[EB/OL]. [2019 – 01 – 09]. http://datacommons.org.nz.
[9] GROSSMAN R L. Data-Commons-Guidelines[EB/OL]. [2019 – 01 – 09]. https://www.healthra.org/wp-content/uploads/2018/08/Data-Commons-Guidelines_Grossman_8_2017.pdf.
[10] FRENCH S P, BARCHERS C V. Designing a data commons for ur-

ban big data[EB/OL]. [2019-01-09]. https://www.rd-alliance.org/final-report-income-streams-data-repositories.html.

[11] VOLCHENBOUM S, HAWKINS D, FRAZIER L, et al. Building pediatric cancer data commons[EB/OL]. [2019-01-09]. https://ascopubs.org/doi/full/10.1200/EDBK_175029.

[12] VOLCHENBOUM S L, COX S M, HEATH A, et al. Data commons to support pediatric cancer research[J]. American Society of Clinical Oncology Educational Book, 2017, 37(24): 746-752.

[13] SANSONE S A, MCQUITON P, ROCCA-SERRA P, et al. FAIR sharing_working_with_and_for_the_community_to_describe_and_link_data_standards_repositories_and_policies[EB/OL]. [2019-01-09]. https://www.researchgate.net/publication/326462185.

[14] BIZER C, MEUSEL R, PRIMPELI A. The web data commons microdata, RDFa and microformat dataset series[EB/OL]. [2019-01-09]. https://link.springer.com/chapter/10.1007/978-3-319-11964-9-18.

[15] PURTOVA N. Health data for common good: defining the boundaries and social dilemmas of data commons[EB/OL]. [2019-01-09]. http://link.springer.com/chapter/10.1007/978-3-319-48342-9-10.

[16] MORGAN M, DAVIS S R. Genomic data commons: a bioconductor interface to the NCI genomic data commons[EB/OL]. [2019-01-09]. https://github.com/seandavi/GenomicDataCommons.

[17] SU Z, BERTAGNOLLI M M, SARTOR A O, et al. A novel, open-access data commons for improved disease management in patients (pts) with Merkel cell carcinoma (MCC)[J]. Journal of clinical oncology, 2018, 36(15): 215-255.

[18] SCOTT C, WALTER S, EDDIE S, et al. VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements[J]. Frontiers in immunology, 2018, 9(39): 976-1002.

[19] HALPHIN P N, READ A J, BEST B D, et al. OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles[J]. Marine ecology progress series, 2006, 77(316): 239-246.

[20] EVANS B J. Barbarians at the gate: consumer-driven health data commons and the transformation of citizen science[J]. American journal of law & medicine, 2016, 42(4): 651-685.

[21] BAMBAUER J Y. Tragedy of the data commons[J]. SSRN electronic journal, 2011, 62(25): 120-135.

[22] 宋秀芬,邓仲华. 基于数据监护的机构知识库研究[J]. 图书馆学研究,2016,31(2):44-48.

[23] 覃丹. 英美社会科学数据管理与共享服务平台调查分析[J]. 图书情报工作,2014,58(8):67-75.

[24] 完颜邓邓. 澳大利亚高校科学数据管理与共享政策研究[J]. 信息资源管理学报,2016(1):30-37.

[25] 杨鹤林. 从数据监护看美国高校图书馆的机构库建设新思路-来自 Data Star 的启示[J]. 大学图书馆学报,2012(2):23-28,73.

[26] 殷沈琴,张计龙,张莹等. 社会科学数据管理服务平台系统选型研究——以复旦大学社会科学数据平台为例[J]. 图书情报工作,2013,57(19):92-96.

[27] 朱玲,聂华,崔海媛,等. 北京大学开放研究数据平台建设:探索与实践[J]. 图书情报工作,2016,60(4):44-51.

[28] 殷沈琴,张计龙,张莹,等. 社会科学数据管理服务平台系统选型研究——以复旦大学社会科学数据平台为例[J]. 图书情报工作,2013,57(19):92-96.

[29] 邓仲华,黄雅婷. “互联网+”环境下我国科学数据共享平台发展研究[J]. 情报理论与实践,2017,40(2):128-132.

[30] 刘兹恒,曾丽莹. 我国高校科研数据管理与共享平台调研与比较分析[J]. 情报资料工作,2017(6):90-95.

[31] 刘桂锋,张裕,刘琼. 科研数据开放平台评价指标体系构建及案例研究[J]. 图书情报知识,2019(1):21-31.

[32] 美国开放数据云联盟[EB/OL]. [2019-03-09]. www.opensciencedatacloud.org.

[33] 复旦大学数据中心[EB/OL]. [2019-03-09]. https://dvn.fudan.edu.cn/home/static/profile.jsp.

[34] 北京航空航天大学数据共享平台[EB/OL]. [2019-03-09]. http://etc.xzit.edu.cn/01/19/c56a281/page.htm.

[35] 清华大学数据共享平台[EB/OL]. [2019-03-09]. http://www.chinaz.com/news/2016/0105/492077.shtml.

[36] 中国科学院计算机网络信息中心[EB/OL]. [2019-03-09]. http://www.nsdata.cn/resource/list? code=1803710.

[37] 中国科学院数据共享平台[EB/OL]. [2019-03-09]. http://www.geodata.cn/.

[38] 武汉大学图书馆数据共享中心[EB/OL]. [2019-03-09]. http://www.lib.whu.edu.cn/kxsj/aboutus.htm.

[39] 华中科技大学科学数据中心[EB/OL]. [2019-03-09]. https://cmis.csdc.info/to about.action.

[40] 清华大学经济社会数据中心[EB/OL]. [2019-03-09]. http://www.sem.tsinghua.edu.cn/sercent/jjshsjzx.html.

[41] 国际神经母细胞瘤风险组. INRG data commons[EB/OL]. [2019-03-09]. http://europepmc.org/abstract/MED/28561664.

[42] HEATH A P, GREENWAY M, POWELL R, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets[J]. Journal of the American Medical Informatics Association, 2014, 21(6): 969-975.

[43] 魏来,高希然. 大数据背景下高校数据馆员的角色定位[J]. 情报资料工作,2015(5):90-94.

[44] MANSELL J, LAKING R, MATHESON B, et al. Data commons blueprint: a high trust, lower cost alternative to enable data integration and reuse[EB/OL]. [2019-03-09]. http://datacommons.org.nz, 2017.

作者贡献说明:

吴雅威:论文选题、撰写和修改;

张向先:论文审阅、修改及最终定稿。

Investigation and Analysis of Foreign Data Commons Platform

Wu Yawei Zhang Xiangxian
School of Management, Jilin University, Changchun 130022

Abstract: [**Purpose/significance**] This paper investigated and analyzed the data management mode of data commons in foreign countries, to promote the research and practice of data management services in China. [**Method/process**] By combing and summarizing the development state of data commons at home and abroad, comparing and analyzing the gap between the two and taking the US-INRG data commons as an example, from the principle and protocol mode, database and user interface and data identification and association. Analyze its data space management model and propose strategies for construction and development of China data commons. [**Result/conclusion**] Combining the case and the state of China's data management platform, the paper puts forward specific suggestions such as overall plan, construction goals, problems to be solved, data commons overall architecture and user service so on.

Keywords: data commons data management data service